MEGA – UE5

In this exercise, you will use **Python** to explore and get insights into a large cohort of COVID patient cases using traditional machine learning approaches for stratification (i.e., dimensionality reduction and clustering). You will work with EHR data (***data_ehr.csv***) that you can find in the **UE5-data.zip** file. There is also a ***data_description.csv*** that includes a description of the variables included in this EHR data set.

For the submission, you need to provide **two deliverables**:

1- A .pdf with the answers to the questions (***a-f***) below.
2- Your .py script(s).

Part 1: Data Preprocessing

Read the EHR data (***data_ehr.csv***), for instance with the **Pandas** library. The dataset contains missing values that are encoded with **NA** or empty values. Clean the dataset by removing columns where more than **30%** of the values are **NA**.

- ***a-*** ***How many dimensions (attributes) and instances (cases or number of patients) does the clean dataset contain? How many dimensions and instances did it contain before the cleaning?***
- ***b-*** ***What type of data does the clean dataset contain?***

Part 2: Clustering

Read the documentation of the **scikit-learn** library for clustering (https://scikit-learn.org/1.5/modules/clustering.html).

Try a couple of different clustering algorithms to cluster the cohort of patients, based on **numerical features** only. Then, decide which algorithm and with what parameters is best to use in this case.

- ***c-*** ***Explain your choices of clustering methods and all parameters used. Why did you use this clustering method and parameters? What are its advantages/disadvantages over other potential methods?***

Part 3: Dimensionality Reduction and Clustering

Read the documentation of the **scikit-learn** library for PCA and t-SNE (https://scikit-learn.org/dev/modules/generated/sklearn.decomposition.PCA.html#sklearn.decomposition.PCA and https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html respectively).

d- What are the advantages/disadvantages of these two dimensionality reduction approaches? Compare the two approaches against each other, theoretically.

Perform these two dimensionality reductions on your dataset and visualize the results as a scatterplot. Subsequently, perform a clustering on top of your dimensionality reduction, according to your observations in Part 2 of this exercise. Show your results again in a scatterplot.

e- What are the advantages of using clustering on top of a dimensionality reduction?

On top of your clustered dimensionality reduced outcome from the previous step, encode in the scatterplot the '2345-7_Glucose [Mass/volume] in Serum or Plasma' variable. For this, you will need to use appropriate visual channels (e.g., opacity, shape, etc.) for the datapoints.

f- Explain the visual variable choices you made in visualizing the glucose variable. Do you have any interesting observations regarding the patterns in the data?